Capitalisation des savoirs par le web : une application de la TEI pour l'encodage et l'exploitation des textes de la Base de Français Médiéval

Serge Heiden, Céline Guillot

Capitalisation des savoirs par le web : une application de la TEI pour l'encodage et l'exploitation des textes de la Base de Français Médiéval

Serge Heiden
Céline Guillot
{slh,cguillot}@ens-lsh.fr

FRE2546 « Analyses de corpus linguistiques, usages et traitements »
CNRS/ENS-LSH, Lyon

Sommaire

Introduction

Les questions relatives à la mise sur le web de corpus d'ancien et de moyen français se sont présentées à nous lors du transfert des textes de la Base de Français Médiéval sur le site français de l'ATILF. Jusqu'à présent, les textes de la Base de Français Médiéval n'étaient pas accessibles sur la toile. Ils étaient exploités par l'intermédiaire de concordances, réalisées au moyen du logiciel ANALYSER de Pascal Bonnefois, mais leur accès restait réservé.
Le transfert de ces textes et leur intégration dans la grande base FRANTEXT nous a donc fourni l'occasion de nous poser un certain nombre de questions méthodologiques. La transmission de données textuelles peut en effet s'envisager selon deux points de vue : on peut échanger des textes entre institutions ou collègues en laissant à chacun le choix des outils grâce auxquels ils les exploiteront, ou l'on peut accéder à des textes que l'on ne possède pas via une base de données interrogeable à distance. C'est le parti pris par les bases FRANTEXT et ARTFL notamment. Dans le cadre de notre accord avec l'ATILF, nous nous situons clairement dans le second cas de figure. Mais la question du mode de représentation des textes ne nous a pas paru secondaire pour autant. Nous présenterons, dans cet article, les modalités de représentation des textes utilisées nous garantissant un contrôle de la qualité des textes

transmis à notre partenaire ainsi que de leur intégration dans notre propre outil d'analyse automatique WEBLEX.

Présentation succincte de la BFM

La Base de Français Médiéval, constituée depuis 1989 sous la direction de Christiane Marchello-Nizia par l'Equipe Linguistique et Informatique puis l'UMR "Analyses de corpus linguistiques, usages et traitements" de l'Ecole Normale Supérieure Lettres et Sciences Humaines[1], rassemble une soixantaine de textes d'ancien et de moyen français. Ces textes appartiennent à des genres variés (roman, poésie, chanson de geste, théâtre, textes juridiques…) et à des époques différentes (de 842 à la fin du 15ème siècle). Ils ont également la particularité d'avoir été saisis dans leur intégralité sur des éditions critiques. La gestion de cette base ainsi que la coordination des relectures sont assurées par Céline Guillot. La normalisation TEI des textes est coordonnée par Serge Heiden.

I. De la représentation des textes

La fréquentation des manuscrits et des textes anciens rend les philologues que nous sommes particulièrement sensibles à la question du mode de présentation du matériau textuel[2]. Ce à quoi nous accédons n'est jamais le texte originel, mais l'une de ses représentations. A cet égard, nous nous élevons contre une conception de l'informatique comme simple moyen de stockage d'une quantité toujours croissante d'informations. Bien au contraire, le mérite des outils numériques est, selon nous, de renvoyer toujours et encore au problème crucial du mode de représentation des textes et de son explicitation. Représenter un texte, cela revient à expliciter les choix d'encodage de l'information qu'il contient. En offrant la possibilité de séparer la *représentation* d'un texte au sens de son stockage sur un support, de sa *présentation* au sens de son affichage sur divers supports (écran, papier, etc) pour consultation et exploitation au sens large, l'informatique a fait entrer l'objet « texte » dans une nouvelle ère.

1.1. Les textes sont à représenter sur un support informatique

Ce cadre théorique une fois posé, l'un des pré requis de l'encodage électronique des textes est qu'il doit être approprié à la machine. De ce point de vue, nous avons voulu choisir une représentation textuelle qui soit accessible à tout programme informatique pour l'exploitation électronique des textes. Nous avons donc écarté les formats dits "propriétaires", comme celui des documents de Word, qui ont la particularité de ne pouvoir être lus que par les logiciels qui les ont créés. Il nous fallait au contraire un mode de représentation que tout logiciel mais aussi toute personne soit susceptible de lire. C'était pour nous une garantie et une assurance qu'il serait toujours possible de faire face à des problèmes imprévus : difficultés liées à l'utilisation d'un logiciel, encodage erroné, etc. C'était également répondre à notre objectif premier de transmettre à nos collègues de Nancy nos textes et nos choix de représentation.

[1] Et grâce notamment à la subvention de l'Institut Universitaire de France.

[2] cf. la préface de Jacques Monfrin à l'ouvrage de Claire Blanche-Benveniste et Colette Jeanjean (1987), *Le français parlé. Transcription et édition*, Didier Erudition, Paris. Ainsi que la section « Où est le texte authentique » pp. 112-115, dans le même ouvrage.

Un autre impératif a motivé notre choix. Nous voulions non seulement transmettre des textes, mais aussi nous donner les moyens de thésauriser le travail interprétatif réalisé sur eux. Il nous semble en effet que la question de l'accumulation du "savoir" interprétatif (linguistique ou autre) est devenu un enjeu de taille. Et là encore le numérique ouvre de nouvelles possibilités, puisqu'il permet d'échanger avec les textes les méta-données y afférant. Une bonne transmission ne peut toutefois se faire qu'à la condition d'être indépendante de l'évolution technologique des systèmes informatiques. La question du mode de représentation choisi et de sa lisibilité se pose donc à nouveau à ce stade, et cela d'autant que l'explicitation des choix d'interprétation rend également possible le contrôle par un tiers de la qualité de l'encodage de ces méta-informations et de leur compatibilité avec d'autres choix.
Ces données (textes et méta-données), qui sont contenues dans un fichier informatique, doivent par ailleurs pouvoir faire l'objet de diverses exploitations informatiques. En d'autres termes, l'accès aux textes est nécessairement "médiatisé" par l'outil informatique. Les modalités d'accès aux textes sont également sensibles à leur localisation. S'ils sont accessibles localement sur les machines qui les interrogent, les possibilités d'exploitation sont a priori plus variées que s'ils sont télé-accessibles. L'une des principales contraintes de l'accès à distance est notamment qu'il peut interdire la transmission du contenu du document dans sa totalité.
On distingue généralement trois grands types d'usage des textes électroniques, ces usages conditionnant aussi leur mode de représentation. Le premier d'entre eux, que nous appelons la lecture, a trait aux différents types d'édition possibles du document : lecture à l'écran, impression sous divers formats ou édition grâce à un traitement de texte. La version électronique du texte doit de ce fait contenir en elle-même (autrement dit, dans son balisage) les informations nécessaires à la création de toutes ces formes éditoriales. Des indications sur la typographie, la mise en page, la pagination, etc. doivent donc être insérées dans le texte ou être calculables à partir de lui.
Les textes peuvent également faire l'objet d'interrogations sur des mots ou des notions. Cette activité de recherche peut elle aussi, et elle a parfois tout intérêt à le faire, s'appuyer sur un balisage explicite au fil du document des notions et/ou des locutions. Les balises servent également à retrouver et à restituer les références bibliographiques attachées aux résultats de la recherche. Il s'agit en général du titre de la section, du numéro de la page et éventuellement de la section, du numéro du paragraphe, du vers…
Le troisième usage possible des textes relève d'avantage d'un travail de synthèse. La comparaison automatique du vocabulaire présent à l'intérieur d'un ou de plusieurs textes est un exemple de calcul de synthèse. Il s'appuie sur une délimitation explicite (au moyen d'un balisage) des différentes parties ou sections d'un texte. Il peut également porter sur plusieurs textes regroupés et délimités dans un même fichier informatique.

1.2. Quel format de représentation utiliser ? (XML)

Notre souci était donc double. Il s'agissait de choisir non seulement un format de représentation de nos textes, mais aussi un format qui rende explicites les méta-informations insérées au fil du texte (les balises). La norme XML offre cette double possibilité puisqu'elle détermine strictement la morphologie des balises qui véhiculent les méta-informations. Dans cette norme par exemple, toute balise commence obligatoirement par un chevron ouvrant "<" et se termine par un chevron fermant ">". Tout ce qui se trouve entre chevrons est donc clairement identifié comme une donnée méta-textuelle destinée à encoder une information portant sur le texte lui-même. L'information méta-textuelle est décrite par le nom de la balise

d'une part, et par l'ensemble de ses propriétés d'autre part. Ces propriétés sont données sous forme de noms d'attributs auxquels sont attachées des valeurs.
La norme XML permet également de représenter la structure du texte. Elle le fait au moyen d'une structure arborescente. On lui reproche parfois de n'être pas adaptée à toutes les œuvres et d'influencer a priori l'interprétation des éléments du texte. L'usage montre au contraire qu'on s'en accommode très bien et qu'elle est le plus souvent bien adaptée à l'encodage des informations méta-textuelles. En tout état de cause, une structure de ce type est toujours plus utile qu'une structure plate (ce qui revient à une absence de structure).

1.3. Quels types d'information faut-il mettre dans les textes et quand ? (TEI)

Comme cela a déjà été dit, la norme de représentation XML permet d'encoder la valeur sémantique des balises dans leur nom et leurs attributs. Les balises peuvent par ailleurs s'emboîter les unes dans les autres à l'intérieur d'une structure arborescente. La place qu'elles occupent dans cette structure, autant que le nom qu'elles portent, fait également partie de leur valeur sémantique. On sait par exemple qu'une balise <p> qui encadre le paragraphe ou la strophe se trouve nécessairement à un niveau inférieur de la balise qui indique une division du type livre, chapitre, etc.
Pour ce qui est de la forme des noms à donner aux balises, nous avons choisi les recommandations de la TEI[3]. Elle offre en effet l'avantage principal de représenter un consensus international sur le sujet. Ce choix doit s'entendre au sens où nous puisons dans la TEI uniquement ce qui nous semble nécessaire à l'exploitation ultérieure des textes par nous-mêmes ou nos partenaires. Nous n'avons jamais envisagé une utilisation exhaustive de cette norme. Elle ne saurait donc être considérée comme une norme d'encodage total des informations d'un texte, mais plutôt comme une méta-terminologie consensuelle dans laquelle il est possible de trouver, en fonction des objectifs d'analyse et de recherche poursuivis, les balises nécessaires à l'encodage des documents. La sémantique de la représentation des textes sous-tendue par la TEI ne doit donc pas être crainte comme un « prêt à penser » influençant l'analyse. Nous la voyons plutôt comme une sorte de zone de forte densité des « marques de pas » jalonnant les divers parcours interprétatifs des textes, cette zone méritant d'être « balisée » par une terminologie partagée par la communauté. D'ailleurs, nous avons d'ores et déjà rencontré les limites de la TEI pour certains textes, ce qui nous amènera peut-être à terme à en proposer une extension[4].
Cet encodage se doit également d'être, selon nous, progressif. Certaines des informations ne pourront être introduites qu'après coup, suite à une analyse. Cela a été notamment le cas dans quatre textes de la Base de Français Médiéval qui ont reçu un enrichissement morpho-syntaxique ultérieur à leur entrée dans la base. Mais surtout, la possibilité de gérer de manière cohérente des encodages intermédiaires permet d'adapter le coût d'encodage à l'usage réel que l'on fait des textes dans un projet donné, un projet ultérieur pouvant enrichir le balisage plus avant si nécessaire. C'est dans cet ordre d'idées que nous avons décidé de ne pas encoder le discours direct dans nos textes pour le projet en cours.
Avant d'entrer plus avant dans le jeu de balises utilisé lors de l'encodage de nos documents, nous souhaiterions préciser les principes qui sous tendent la notion de texte dans notre base.

[3] Voir http://www.tei-c.org . La TEI n'est pas à proprement parler une norme, au titre de l'ANSI, l'AFNOR ou de l'ISO par exemple, mais un consortium proposant un travail éditorial annuel sur les recommandations qu'elle produit. C'est un mode de fonctionnement qui se rapproche de celui du W3C pour les « normes » HTML ou XML par exemple.
[4] La notion d'extension est d'ailleurs déjà prévue dans le cadre de la TEI.

1.4. Qu'est-ce qu'un texte ?

Sur le plan informatique, l'unité textuelle est par définition le fichier numérique. Ce qui reste à définir en revanche, c'est le contenu de ce fichier. Nous avons pour notre part adopté le principe[5] selon lequel un volume physique correspond à un fichier informatique. Ce principe a le mérite d'être clair. Il conduit néanmoins dans un certain nombre de cas à des situations sinon aberrantes, du moins un peu étranges. *Le Roman de Thebes* par exemple, qui comporte deux volumes, est traité comme deux unités textuelles. Les *Lais* de Marie de France au contraire constituent un seul texte. Il faut cependant limiter ces considérations de fichiers aux simples contingences informatiques. La notion d'œuvre prendra réellement son sens dans les outils d'exploitation de la base de textes qui, par exemple, seront capables d'interpréter la structuration d'un fichier avec ses balises pour y découvrir les différents lais au besoin.
La question de la définition de l'unité textuelle une fois résolue, il reste encore à faire clairement la part, comme le requiert d'ailleurs la TEI, entre le texte en tant que tel et son en-tête. Le texte en lui même peut se définir comme le corps ou le contenu, l'en-tête comme la tête du document contenant des méta-données sur son contenu. On encode et on enregistre dans l'en-tête les informations qui portent sur la totalité du corps : les sources bibliographiques, les principes d'annotation mis en œuvre (en d'autres termes, l'usage particulier de la TEI qu'on trouve dans le texte), l'historique des révisions d'encodage, les objectifs du projet à l'origine de l'annotation informatique, etc. Comme nous le verrons à la section IV, les informations bibliographiques des en-têtes sont gérées dans une base de données rassemblant toutes les fiches descriptives de nos textes.

II. Annotation TEI et vérification de nos textes

2.1. Nos choix d'annotation TEI

Un document réalisé par notre équipe et intitulé "Manuel d'encodage" décrit de manière assez précise nos principes d'encodage et la liste des balises que nous avons choisies parmi toutes celles que propose la TEI. Ce document est une référence aussi bien pour nous que pour ceux à qui nous transmettrons nos textes. Il est également un gage que tous nos textes sont balisés selon des choix et des principes homogènes (il a servi de référence aux personnes qui ont balisé nos textes). Ce document ne doit pas seulement être interprété comme le choix d'un sous-ensemble de la TEI, mais surtout comme la mémoire de nos choix d'encodage, la TEI ayant simplement servi de terminologie permettant, entre autre, l'échange avec nos partenaires. Nous présentons ce document de travail, et par conséquent en constante évolution, en annexe du présent article (Annexe I).

2.2. Outils d'enrichissement et de vérification de l'encodage TEI

Pour nous assister dans le travail d'encodage, une boîte à outils SGML/XML, appelée la Toolkit LML[6] et conçue à partir de la librairie LT XML 1.0 du Language Technology Group d'Édimbourg[7], a été mise au point par Serge Heiden.

[5] Notons que c'est également le choix, dans une certaine mesure, de la base Frantext.
[6] La Toolkit a été développée dans les environnements Unix Solaris et Windows 2000.

La librairie choisie présente l'intérêt fondamental de rendre possible le traitement de documents XML de taille quelconque, en appliquant successivement divers traitements d'enrichissement en flux et en garantissant en permanence la conformité de la structure du document avec sa DTD[8]. Cette librairie permet de manipuler les balises des documents XML via un parcours arborescent (modèle DOM) ou sous la forme événementielle d'un flux d'éléments XML (modèle SAX).
La toolkit LML permet d'effectuer trois types d'opérations fondamentaux sur les textes : leur vérification, leur enrichissement et l'extraction de données.
Vérifier un texte revient tout d'abord à contrôler qu'il est en conformité avec la DTD TEI. Ce contrôle se fait en SGML avec l'outil `sp` de James Clark[9] et en XML avec l'outil `rxp`[10]. Par ailleurs l'outil de la toolkit LML appelé `lmlpp` permet une vérification de la structure du texte en affichant sous une forme indentée la totalité ou une partie de son arborescence XML. Cet affichage combiné avec des outils Unix élémentaires[11] de sélection (`grep`), de tri (`sort`) et de filtrage (`uniq`) permettent, par exemple, de vérifier la conformité du début et de la fin de la numérotation des sections, des pages, des vers…, ainsi que des contrôles croisés du type : numéro du dernier vers de la dernière page ou encore numéro de la page où figure le titre de chaque section…
L'enrichissement des textes concerne d'une part la numérotation automatique de divers éléments (pages, vers, etc.), la projection d'informations sous forme d'attributs affectés à certains éléments (on projette par exemple un attribut type sur chaque réplique d'un personnage dans une pièce de théâtre à partir d'une liste de correspondances associant à chaque nom de personnage un type), la possibilité de transmettre les informations portées par différents attributs d'éléments sur d'autres éléments situés en deçà dans l'arborescence (outil `lmlheritatt`). On peut ainsi associer à un vers son numéro de page en cas de besoin par exemple.
Les outils *d'extraction* permettent enfin de sélectionner n'importe quelle sous arborescence d'un document en utilisant une expression de recherche écrite sous une forme simplifiée de la norme XPATH[12] (outil `lmlget`). Ils permettent aussi (outil `lmlcost`) un traitement final sur les textes. On peut ainsi par exemple produire des éditions HTML à façon à partir de l'arborescence XML-TEI à l'aide d'instructions de transcodage.

III. Nos besoins internes d'exploitation des textes avec l'outil Weblex

3.1. La linguistique de corpus avec l'outil Weblex

La philosophie adoptée par notre équipe étant que seuls les éléments utiles à l'exploitation des textes méritent un encodage, il est bon de présenter brièvement ici notre principal outil d'analyse linguistique. Le logiciel Weblex est né de la démarche expérimentale dite analyse

[7] Dont l'adresse Web est http://www.ltg.ed.ac.uk/ .
[8] Définition de Type de structure de Document. Déclaration formelle des différentes balises et de leurs imbrications. Techniquement, la TEI correspond à une DTD à laquelle les documents encodés en XML doivent se conformer.
[9] Voir http://www.jclark.com/sp/index.htm .
[10] Voir http://www.ltg.ed.ac.uk/~richard/rxp.html .
[11] Adaptés à Windows, voir http://unxutils.sourceforge.net/ .
[12] Voir http://www.w3.org/TR/xpath .

automatique de corpus. Cet outil permet d'appréhender dans nos textes, grâce à l'encodage XML-TEI dont ils sont pourvus, un certain nombre d'observables. Sur la base de ces observables il peut alors produire tout un ensemble de synthèses : de vocabulaire, de collocations, de spécificités d'apparition dans telle ou telle partie d'un texte, de répartitions, de n-grammes, etc. Il peut bien sûr, plus classiquement, servir d'outil d'exploration assistée des textes, grâce en particulier à ses concordances kwic réalisées au moyen d'expressions de recherche pouvant porter sur une succession de plusieurs lexèmes à la fois, éventuellement discontinue, et leurs propriétés (partie du discours, lemme, etc). Ces concordances offrent par ailleurs l'avantage de permettre à l'utilisateur d'ordonner les contextes à sa guise. Toutes les fonctionnalités du logiciel sont accessibles exclusivement par le Web. Ce qui donne potentiellement aux corpus qu'il manipule, des caractéristiques d'accès similaires à ceux des bases Frantext ou ARTFL.

3.2. Intégration d'un corpus de textes dans Weblex

Pour que l'outil Weblex puisse travailler sur un corpus, il faut d'abord procéder à son intégration dans les bases de l'outil. Cette intégration consiste fondamentalement à identifier dans les textes les informations nécessaires aux traitements qui vont suivre et à optimiser leurs accès ultérieurs, notamment en construisant des indexes de mots (ou lexèmes), de propriétés de mots ou encore de structures : paragraphes, phrases, etc.
Il y a quatre groupes d'informations fondamentaux à extraire des textes au moment de leur intégration :

1) la délimitation des lexèmes et la description de leurs propriétés (partie du discours, lemme, etc.) ;
2) la délimitation de la macro-structure logique du texte : parties / sections / paragraphes / propositions-phrases, etc. ;
3) les indications bibliographiques nécessaires à la construction des références de concordances : auteur, date, numéro de page, de vers, etc. ;
4) la délimitation des différentes partitions du corpus qui seront à l'origine des calculs contrastifs : partition en dates pour les analyses diachroniques, en genre pour les analyses génériques, etc.

La figure 1 (page suivante) illustre le procédé employé pour extraire toutes ces informations des textes à l'aide des outils d'extraction d'informations à partir des balises XML-TEI.

Actuellement, l'intégration d'un corpus dans Weblex doit être réalisée par nos soins. Un portail[13] d'accès à l'outil permettra à terme aux utilisateurs d'alimenter de façon autonome les bases de textes et de corpus.

[13] Voir https://weblex.ens-lsh.fr/ .

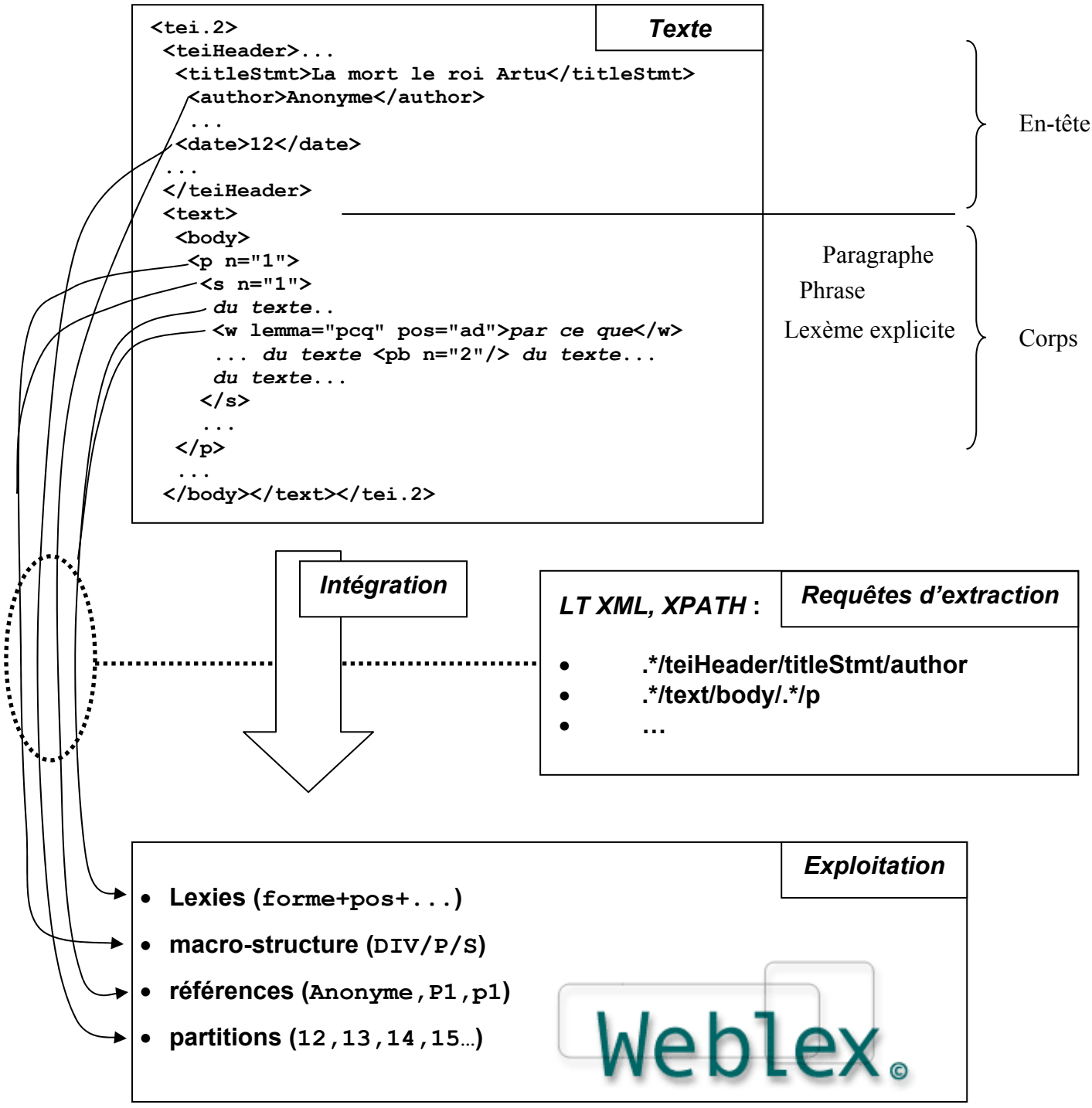

Figure 1.
Schéma synoptique de la procédure d'intégration d'un texte XML-TEI dans les bases d'indexation de l'outil Weblex.

IV. Base de données bibliographiques de la BFM

La Base de Français Médiéval s'organise autour d'une base de données bibliographiques. Certaines des données qui s'y trouvent sont également celles qui figurent dans les en-têtes TEI des différents textes. Un exemple en est donné dans l'Annexe II. D'autres informations sont de nature plutôt pratique (nom du fichier informatique contenant le texte XML-TEI par exemple). D'autres encore ont trait à la tradition manuscrite, à la création et à la réception des textes… Ces informations, trop spécialisées pour trouver place dans le cadre standard de la TEI, sont ici conservées. Les différents champs de la base bibliographique étant par ailleurs indexés, toutes ces données peuvent également faire l'objet de recherches pour trouver des textes. Cette base est implémentée dans un SGBDR (Système de Gestion de Base de Données Relationnelles) interrogeable avec le langage SQL.

Conclusion et présentation des documents annexes

Ce travail d'encodage sur les textes de la BFM repose sur un ensemble de documents de travail :

- le *Manuel d'encodage des textes TEI pour la BFM* (texte de référence pour l'encodage) ;
- les *Consignes pour le balisage des textes de la base de français médiéval* (directives données aux relecteurs/baliseurs) ;
- le *Questionnaire sur l'usage des marques typographiques dans les éditions des textes de la BFM* (document que doivent remplir les relecteurs/baliseurs) ;
- les *Listes d'autorité pour les valeurs d'attributs* (listes de référence) ;
- les *Étapes de vérification finale des textes*[14] (procédures finales de vérification des textes encodés).

Ces documents ont une double fonction : ils constituent à la fois la mémoire de nos choix de représentation et fournissent un guide destiné aux diverses personnes responsables de l'encodage. Ces personnes – une vingtaine au total - ayant des compétences variées, leurs tâches dans le projet varient également. Certaines sont chargées de la relecture du texte[15] et d'un premier balisage (les documents qui les concernent sont les *Consignes* et le *Questionnaire*), d'autres sont responsables des opérations de numérotation automatique et de divers types de projection ainsi que de la vérification finale des textes encodés (*Manuel d'encodage*, *Listes d'autorité* et *Étapes de vérification*).
Ces documents étant spécialisés en fonction de tâches précises, ils peuvent parfois sembler se contredire et/ou se répéter. En réalité, chaque tâche correspondant à un moment du processus d'encodage, les documents qui y sont associés ne reflètent qu'un état intermédiaire d'un texte. De ce point de vue, le *Manuel d'encodage* est le seul document de référence associé à l'étape ultime de l'encodage.
Le fait par ailleurs que ces différents documents aient plusieurs fonctions (texte de référence, ensemble de directives d'encodage, renseignements sur un texte nouveau et procédure de contrôle strict final) provoque un certain nombre de redites. Mais ce phénomène est inhérent au choix que nous avons fait d'adapter les documents aux personnes et aux tâches qui leur sont confiées.

[14] Pour des raisons de place nous n'intègrerons pas ce document, qui est par ailleurs très technique, dans le présent article.
[15] On entend par relecture la simple vérification du texte dépourvu de balises.

Enfin, les documents que nous présentons en annexe sont en permanente évolution. Typiquement, le *Questionnaire sur l'usage des marques typographiques dans les éditions des textes de la BFM* peut servir à signaler l'utilisation de nouvelles marques typographiques absentes des listes d'autorité en cours, ce qui nous conduira alors à mettre à jour ces listes.
Le caractère également progressif de nos pratiques d'encodage est un dernier facteur de variation de l'ensemble de ces documents.

En conclusion, il nous a paru utile de présenter en annexe ces documents, qui ne sont pas une simple reformulation des recommandations de la TEI, mais qui reflètent une pratique effective d'encodage collectif et progressif.

Manuel d'encodage des textes TEI pour la BFM

Normes utilisées à la date du 3 avril 2003

Serge Heiden (slh@ens-lsh.fr)
Céline Guillot (cguillot@ens-lsh.fr)
Alexei Lavrentiev (alexei-lavrentev@mail.ru)

CNRS / ENS-LSH, FRE 2546

(VERSION 1.0- JUILLET 2002)
(VERSION 2.0- JANVIER 2003)

SOMMAIRE

Introduction

Ce document présente l'ensemble des méta-informations susceptibles d'être intégrées dans un texte en vue de sa gestion dans la base de textes et de son traitement automatique par Weblex. Ces méta-informations sont représentées explicitement dans le texte sous la forme de balises SGML/XML[16]. Le nom et la structuration de ces balises correspondent exactement à un sous-ensemble des recommandations de la TEI.
En dehors des balises délimitant le début et la fin du texte **aucune balise n'est obligatoire**.
Le relecteur/encodeur[17] ajoute les seules méta-informations dont il dispose et qu'il croit nécessaires à l'exploitation future du texte. Ce document ne présente donc pas un « format » de texte particulier mais bien un moyen de communication formel entre le chercheur, ses partenaires et les outils d'exploitation permettant d'*expliciter* les notions du texte nécessaires à leur traitement (la structure interne du texte, la mise en page de l'édition de référence, les interventions éditoriales…). La représentation de ces notions se fait au moyen de balises insérées dans le texte.
Pour un encodage plus complet on pourra consulter la documentation de la TEI située à l'url http://www.tei-c.org ou encore les recommandations de la TEI dans son ensemble, à l'url http://www.tei-c.org/Guidelines2/index.html. De manière générale il faudra consulter les TEI Guidelines comme complément aux imprécisions de ce manuel.

Généralités

Principes de base

- les documents à baliser sont au format texte brut (il n'y a pas de police ou de style particulier à utiliser) ;
- un document est composé d'éléments, qui sont délimités par des balises ;
- chaque balise est délimitée par des chevrons (**<, >**) ;
- on distingue les éléments qui contiennent un autre type d'élément (ex : le chapitre contient une portion du texte et il est susceptible d'être divisé en paragraphes...), et ceux qui indiquent uniquement une frontière (ex : le saut de ligne marque la frontière entre deux lignes) :
    - dans le premier cas, la balise qui a un contenu (ou une portée) se place au début de son contenu et le termine par une balise fermante qui possède un caractère « / » en préfixe (ex : ** **...** **) ;
    - dans le second cas, la balise qui n'a pas de contenu est unique (ex : **<pb>**).

Exemple :

```
<div type="chapitre" n="1">
contenu du premier chapitre
</div>
```

[16] Pour faciliter le travail des relecteurs/encodeurs, le balisage se fait pour l'instant d'abord en SGML. Ce codage est ensuite traduit en XML.
[17] La personne qui vérifie la conformité du texte avec l'édition de référence et qui ajoute les balises au texte.

Glose : on appelle élément la partie du document qui commence avec la balise ouvrante ** ** et se termine avec la balise fermante ** **

- toute balise peut posséder plusieurs propriétés en plus de son nom, sous la forme d'une succession de relations *nom-attribut*= "*valeur-attribut*" situées entre les chevrons (ex : **<pb n="2">**, où l'attribut **n** encode le numéro de la page qui débute) ;
- l'ordre dans lequel on indique les attributs est libre ;
- on indiquera le nom et la valeur de l'attribut si on possède cette information ;
- le nombre d'espaces ou de tabulations situés entre deux mots ou balises dans le corps du texte et dans les valeurs d'attributs n'a pas d'interprétation particulière ;
- de même pour les sauts de ligne et leur nombre ;
- les textes doivent être enregistrés en format texte brut (avec un encodage des caractères Windows).

Rôle des différents acteurs participant à l'édition du texte informatisé

Le point de départ de la constitution d'un document étant une édition de référence imprimée, il est nécessaire d'indiquer au fil du texte les interventions de l'éditeur scientifique mises en évidence par des procédés typographiques. Pour passer du document imprimé à la version électronique du texte, trois étapes se succèdent :

1) la numérisation du texte de l'édition de référence ;
2) la vérification que cette version est conforme au texte de l'édition de référence, et le prébalisage du texte ;
3) l'enrichissement du balisage, la vérification formelle de la structure du document SGML/XML et la validation finale.

La TEI prévoit la description formelle de chaque personne intervenant dans la constitution du document. Chaque correction, intervention éditoriale ou commentaire doit porter la mention de la personne qui en est responsable. On utilise pour ce faire l'attribut **resp** auquel on adjoint l'une des quatre valeurs suivantes :

- **"editor"** qui correspond à l'éditeur scientifique ;
- **"numerizor"** qui correspond à la personne qui numérise le texte de l'édition de référence ;
- **"proofreader"** qui correspond à la personne qui relit et fait le prébalisage du texte (si plusieurs personnes sont chargées de ce travail on leur attribue un numéro, exemple : "proofreader1", etc.) ;
- **"encoder"** qui correspond à la personne qui réalise la vérification formelle et la validation finale.

Description formelle du balisage des textes de la BFM

Délimitation du corps du texte et des éléments qui lui sont externes (prologue…)

- l'élément **<text>** regroupe à la fois le corps du texte et tous les éléments qui lui sont externes
- l'élément **<body>** marque le début et la fin du corps du texte proprement dit
- l'élément **<front>** marque les informations liminaires : prologue, sommaire…, et surtout le titre de l'œuvre

- l'élément **<back>** marque les informations supplémentaires : appendice, index…, et surtout l'explicit de l'oeuvre

NOTA BENE :
- en plus de son encadrement par l'élément **<body>**, le contenu du texte doit toujours se trouver dans au moins un élément de structuration, par exemple dans un élément paragraphe **<p>** ;
- dans le même ordre d'idées un élément **<front>** devra toujours encadrer au moins un paragraphe (l'élément **<p>**) et une division ** **, cette division contenant éventuellement un **<head>**
- même chose pour l'élément **<back>**
  - dans le cas où le **<front>** encadre le titre, la balise ** ** a un attribut **type="titre"** et la balise **<p>** n'a pas d'attribut
  - dans le cas d'un explicit, la balise ** ** a un attribut **type="explicit"** et la balise **<p>** n'a pas d'attribut

La structure potentielle d'un texte est donc la suivante :

```
<text>
<front>
<div type="titre">
<p>
titre
</p>
</div>
</front>
<body>
contenu du texte
</body>
<back>
<div type="explicit">
<p>
explicit
</p>
</div>
</back>
</text>
```

Seuls les éléments **<text>** et **<body>** sont obligatoires dans un texte encodé en TEI.

Exemple :

```
<text>
<body>
<p>
Buona pulcella fut Eulalia,
Bel auret corps, bellezour anima.
Uoldrent la ueintre li Deo inimi,
```

```
Uoldrent la faire diaule seruir.
Elle no.nt eskoltet les mals conselliers,
Qu'elle Deo raneiet chi maent sus en ciel.
…
Tuit oram que por nos degnet preier
Qued auuisset de nos Christus mercit
Post la mort et a lui nos laist uenir
Par souue clementia.
</p>
</body>
</text>
```

Structure du corps du texte

Délimitation des parties du texte : livres, chapitres, section, sous-section…

- l'élément ** ** marque n'importe quelle division du texte ;
- des éléments ** ** peuvent contenir d'autres éléments ** ** d'un niveau de structuration inférieur ;
  - l'élément ** ** peut être renseigné avec l'attribut **type** pour indiquer le type de la division (chapitre, section…) et par un attribut **n** pour indiquer son numéro éventuel.

Exemple :

```
…
- Et pour vous informer du temps dont ay eu congnoissance
dudit seigneur, dont faictes demande, m'est force de
commancer avant le temps que je veinse en son service; et
puis, par ordre, je suyvray mon propos jusques à l'heure
quce je devins son serviteur, et continueray jusques à son
trespas.
<div type="livre" n="1">
<div type="chapitre" n="1">
Au saillir de mon enfance et en l'aage de povoir monter
à cheval, fus amené à Lisle devers le duc Charles de
Bourgoigne, lors appellé conte de Charroloys, lequel me
print en son service, et fut l'an mil quatre cens soixante
quatre
…
</div>
</div>
```

Notation des titres de livre, chapitre…

- chaque élément ** ** peut s'ouvrir avec un premier élément **<head>** contenant le titre ou l'entête de la division et se clore par un élément **<trailer>** contenant des informations présentes en fin de division.

Exemple :

```
<div type="chapitre" n="1">
<head>
Ci commence li premiers chapitres qui parole de l'office as baillis.
</head>
Tout soit il ainsi qu'il n'ait pas en nous toutes les
graces qui doivent estre en homme qui s'entremet de baillie,
pour ce ne lerons nous pas a traitier premiers en cest
chapitre de l'estat et de l'office as baillis, et dirons briement
une partie des vertus qu'il doivent avoir, et comment il se
doivent maintenir, si que cil qui s'entremetront de l'office
i puissent prendre aucune essample...
</div>
```

NOTA BENE :

Il peut arriver qu'on rencontre dans un texte un titre qui ne correspond à aucune division logique (et qui n'est pas numéroté). On considèrera alors qu'on a affaire à une pseudo-division qu'on balisera au moyen de la balise ** **…** **. Le titre se trouvera comme dans les cas précédents dans son **<head>**.

Numérotation des pages

- l'élément **<pb>** marque les sauts de page :
  - il a un attribut **n** qui permet d'encoder le numéro de la page qui s'ouvre ;
  - il a aussi un attribut **ed** qui permet éventuellement de préciser de quelle édition provient la pagination (en cas d'annotation simultanée de la pagination de plusieurs éditions).

Exemple :

```
…
- Or i voist donc, fait ele, car se il demain ne deust
revenir il n'i alast hui par ma volenté.» Et il monte
et la damoisele ausi, <pb n="2"> si se partent de laienz sanz
autre congié, et sanz plus de compaignie, fors
solement dui escuier qui avec la damoisele
estoient venuz. Et quant il sont issu de Kamaalot
…
```

NOTA BENE :

Il arrive que des illustrations s'intercalent dans le texte. Si elles occupent la totalité d'une page, il est nécessaire d'insérer un saut de page dans le texte à l'endroit où elles se trouvent (y compris si elles ne sont pas conservées dans la version numérique du texte).

Délimitation des unités inférieures

Textes en prose

- l'élément **<p>** marque les paragraphes ;
- l'élément **<lb>** marque les sauts de ligne.

Exemple :

```
<p>
Or i voist donc, fait ele, car se il demain ne deust<lb>
revenir il n'i alast hui par ma volenté.» Et il monte<lb>
<pb n="2">
et la damoisele ausi, si se partent de laienz sanz<lb>
autre congié, et sanz plus de compaignie, fors<lb>
solement dui escuier qui avec la damoisele<lb>
estoient venuz. Et quant il sont issu de Kamaalot<lb>
</p>
```

Textes en vers

- l'élément **<p>** marque les groupes de vers : laisses, strophes, refrains…
  - il a obligatoirement un attribut **rend** auquel on peut donner plusieurs valeurs : laisse, strophe… Si la qualification des groupes de vers d'un ouvrage est difficile, l'attribut **rend** a la valeur **"gv"** (groupe de vers)[18].
- l'élément **<lb>** marque les fins de vers, y compris quand le vers est incomplet.

Exemple :

```
<text>
<body>
<p rend="strophe">
Buona pulcella fut Eulalia,<lb>
Bel auret corps, bellezour anima.<lb>
…
Post la mort et a lui nos laist uenir<lb>
Par souue clementia.<lb>
</p>
</body>
</text>
```

[18] Il nous est impossible d'utiliser dans ce cas l'élément **<lg>** recommandé par la TEI, parce qu'il impose qu'on balise chaque vers au moyen de l'élément **<l>** qui entraîne par ailleurs la présence de contraintes indésirables.

Théâtre

- l'élément **<sp>** marque les prises de parole :
  - son attribut **who** permet de renseigner le nom du locuteur ;
- l'élément **<stage>** marque les didascalies.

Exemple :

```
<sp who="Pathelin">
Encor ne le dis je pas pour me
vanter, mais n' a, au territoire
ou nous tenons nostre auditoire,
homme plus saige fors le maire.
</sp>
<sp who="Guillemette">
Aussy a il leu le grimaire
et aprins a clerc longue piece.
</sp>
```

Indications dans le corps du texte

Corrections et interventions éditoriales

Les éditions de textes en ancien français contiennent souvent des marques typographiques (italiques, caractères gras, majuscules, crochets, etc.) qui servent à mettre en évidence des passages dans une langue étrangère, un changement de manuscrit, une coquille, une intervention éditoriale, etc. Les pratiques varient selon les éditions, et il convient donc d'analyser leur usage dans chaque édition et de baliser ces passages conformément à leur nature avec les marques qu'offre la TEI.

Corrections de l'éditeur scientifique

- l'élément **<corr>** sert à marquer les corrections effectuées par l'éditeur scientifique et qui sont indiquées de façon explicite dans le corps du texte de l'édition ;
  - son attribut **resp** indique le responsable de la correction, c'est-à-dire l'éditeur scientifique ("editor") ;
  - son attibut **sic** permet d'indiquer le texte original remplacé par la correction. Il sera vide (**sic=""**) en cas d'ajout (en cas, par exemple, d'une préposition manquante). Il sera absent si l'éditeur scientifique n'indique pas le texte original sur la même page ;
  - son attribut **rend** permet d'indiquer la marque typographique utilisée dans l'édition pour mettre la correction en évidence. Il convient de l'utiliser si l'usage de cette marque typographique n'est pas homogène au sein de l'édition ;
  - son attribut **cert** permet d'indiquer éventuellement la certitude de la correction.

Exemple :

```
Un suen humme i out mis pur le lit <corr rend="[a]" sic=""
resp="editor">a</corr> guarder.<lb n="1997">
```

Glose : Ici, d'après l'éditeur scientifique, une préposition *a* "manque dans tous les manuscrits", il l'a donc ajoutée entre crochets. Ces crochets peuvent être trouvés dans l'attribut **rend**, et l'élément **sic** indique qu'il s'agit d'un ajout par rapport à la source de l'édition.

Propositions de corrections du relecteur/encodeur

- L'élément **<sic>** peut être utilisé par le relecteur ou l'encodeur pour proposer une correction d'une erreur flagrante de l'édition. La règle fondamentale de la représentation numérique des textes de la BFM est que le texte numérisé doit être la copie exacte du texte papier de l'édition de référence. La correction est proposée dans un attribut de la balise, et le contenu de l'élément reste une reproduction fidèle de l'édition de référence Le texte de l'édition **ne saurait donc être modifié** ;
  - l'attribut **resp** indique quelle est la personne responsable de la correction (**"proofreader"** ou **"encodeur"**) ;
  - l'attibut **corr** contient la correction proposée. Il sera vide (**corr=""**) en cas de suppression (en cas, par exemple, d'un mot répété), et le contenu de l'élément sera vide en cas d'ajout ;
  - l'attribut **cert** permet éventuellement au relecteur d'indiquer s'il est sûr ou non de la correction(**cert="yes"** ou **"no"**) ;
  - le relecteur peut utiliser l'élément **<note>** pour ajouter un commentaire (cf. 3.3 ci-dessous).

Exemple :

```
<sic resp="encoder" corr="Ço">Co</sic> que li plus halz fist plus
bas peüst desfaire;<lb n="4904">
```

Glose : Ici le démonstratif est noté avec *C* sans cédille dans l'édition. Il y a cependant d'autres occurrences de ce même pronom dans l'édition où la cédille est utilisée. L'encodeur, qui a constaté cette incohérence, a donc décidé de la corriger en mentionnant la graphie de l'édition dans l'attribut **sic**.

Lacunes du manuscrit indiquées par l'éditeur scientifique

- l'élément **<gap>** permet d'indiquer des lacunes dans le texte constatées par l'éditeur scientifique ;
  - son attribut **resp** permet d'indiquer la personne qui a constaté la lacune, c'est-à-dire l'éditeur scientifique (**resp="editor"**) ;
  - son attribut **rend** permet d'indiquer éventuellement la marque typographique utilisée dans l'édition (**"…"**, par exemple).

- son attibut **desc** permet éventuellement de décrire la nature de la lacune (par exemple, **"manuscrit endommagé"**, **"vers omis"**). Il convient de se référer aux notes de l'éditeur scientifique pour renseigner cet attribut ;
- son attribut **extent** permet éventuellement d'indiquer l'ordre de grandeur de la lacune. On peut utiliser des valeurs comme **"1 line"**, **"0,5 line"**, etc.

Exemples :

```
L'autres gaians, qui rostissoit<lb n="719">
<gap resp="editor"><lb n="720">
Et aveuc son poivre faisoit.<lb n="721">
```

```
Dedens vont, regardent les <gap resp="editor"><lb n="6068">
Afaitent les, metent
```

Passages difficiles à lire dans le manuscrit

- l'élément **<unclear>** permet de marquer des passages qui ne sont pas clairs dans le manuscrit source et que l'éditeur scientifique met en évidence à l'aide de marques typographiques. Normalement, son usage ne concerne que les éditions diplomatiques ;
  - son attribut **resp** sert à indiquer la personne qui a mis en évidence le passage incertain, c'est-à-dire l'éditeur scientifique (**resp="editor"**) ;
  - son attribut **rend** permet éventuellement d'indiquer la marque typographique utilisée dans l'édition. Il convient de l'utiliser si l'usage de cette marque typographique n'est pas homogène au sein de l'édition ;
  - son attribut **reason** permet d'indiquer éventuellement la raison pour laquelle le passage est considéré comme n'étant pas clair (par exemple, **"illegible"** ou **"ambiguous"**) ;

Exemple :

```
Que les prisons touz uos r<unclear resp="editor">en</unclear>drai. <lb>
```

Autres types de passages mis en évidence

Passages en langue étrangère

- L'élément **<foreign>** marque les passages écrits dans une langue différente de celle du texte.
  - Si c'est du latin, ce qui est le cas le plus fréquent, la balise **<foreign>** est suffisante. Si c'est une autre langue, il convient d'ajouter l'attribut **lang** dont la valeur précise quelle est la langue utilisée ;
  - l'attribut **rend** permet d'indiquer la marque typographique employée. Il convient de l'utiliser uniquement si l'usage de la marque typographique n'est pas homogène au sein de l'édition.

Exemple :

```
…
- Et tout li haut
homme, et clerc et lai et petit et grant, demenerent
si grant goie a l'esmovoir que onques encore si faite
goie ne si fais estoires ne fu veus ne oïs; et si fisent
li pelerin monter as castiaus des nes tous les
prestres et les clers qui canterent <foreign> Veni creator spiritus
</foreign>
Et trestout et grant et petit plorerent de pec et de
le grant goie qu'i eurent
```

Mises en évidence typographiques dont la signification n'est pas claire

- L'élément **<hi>** marque les passages imprimés dans une typographie différente de celle qu'on trouve habituellement dans le texte et dont on n'a pas d'interprétation particulière ;
  - la marque typographique de l'italique dans un texte est à indiquer à l'aide d'un élément **<hi rend="ital">…</hi>** ;
  - la marque typographique de mots en majuscules dans un texte est à indiquer à l'aide d'un élément **<hi rend="maj">…</hi>** ;
  - la marque typographique de mots en petites majuscules dans un texte est à indiquer à l'aide d'un élément **<hi rend="pmaj">…</hi>** ;
  - la marque typographique de mots en exposant dans un texte est à indiquer à l'aide d'un élément **<hi rend="exp">…</hi>**.

Exemple :
jj^{c} → jj**<hi rend="exp">**c**</hi>** ;

  - la marque typographique de mots en indice dans un texte est à indiquer à l'aide d'un élément **<hi rend="ind">…</hi>**.

Notes et commentaires du relecteur/encodeur

Nous avons fait le choix de ne pas conserver dans notre version numérique du texte les notes de l'éditeur scientifique (dans lesquelles en particulier il indique d'autres variantes du texte).

- l'élément **<note>** marque les annotations et les commentaires du relecteur/encodeur.
  - son attribut **resp** doit indiquer quelle est la personne responsable de la note : le relecteur (**"proofreader"**), l'encodeur (**"encoder"**)…

Exemple :

```
A Com cist cheualiers qui ci siet. <note resp="proofreader">Deux
lettres majuscules initiales</note> <lb n="423">
Qu'il ne respont ne un neel.
```

Glose : l'éditeur indique la présence de 2 majuscules en début de ligne.

Index des balises et des attributs


```
<!DOCTYPE TEI.2 PUBLIC "-//TEI P4//DTD Main DTD Driver File//EN"
"/dtd/sgml/tei/tei2.dtd" [
<!ENTITY % TEI.general "INCLUDE" >
<!ENTITY % TEI.corpus "INCLUDE" >
<!ENTITY % TEI.analysis "INCLUDE" >
<!ENTITY % TEI.certainty "INCLUDE" >
<!ENTITY % TEI.figures "INCLUDE" >
<!ENTITY % TEI.linking "INCLUDE" >
]>
<!-- En-tête non renseigné -->
<tei.2>
<teiheader type='text' status='new'>
<filedesc>
<titlestmt>
<title></title>
</titlestmt>
<publicationstmt><p></p></publicationstmt>
<sourcedesc default='no'>
<p><name id='latin' type='lang'></name></p>
</sourcedesc>
</filedesc>
</teiheader>
<text>
<front>
<div type="prologue">
<head>Prologue</head>
<pb n="1">
<p>
Monsieur l'arcevesque de Vienne, pour satisfaire à la<lb>
requeste qu'il vous a pleu me faire de vous escrire et mettre<lb>
par memoire ce que j'ay sceu et congneu des faictz du roy<lb>
Loys unziesme, à qui Dieu face pardon, nostre maistre et<lb>
bienfaicteur, et prince digne de très excellente memoire, je<lb>
```

```
l'ay faict le plus près de la verité que j'ay peu et sceu avoir<lb>
souvenance…<lb>
</p>
<p>
…Et pour vous informer du temps dont ay eu congnoissance<lb>
dudit seigneur, dont faictes demande, m'est force de<lb>
commancer avant le temps que je veinse en son service; et<lb>
puis, par ordre, je suyvray mon propos jusques à l'heure<lb>
quce je devins son serviteur, et continueray jusques à son<lb>
trespas.<lb>
</p>
<pb n="2">
</div>
</front>
<body>
<div type="livre" n="1">
<div type="chapitre" n="1">
<head>Débuts de Commynes au service de Charles le Téméraire</head>
<p>
Au saillir de mon enfance et en l'aage de povoir monter<lb>
à cheval, fus amené à Lisle devers le duc Charles de<lb>
Bourgoigne, lors appellé conte de Charroloys, lequel me<lb>
print en son service, et fut l'an mil quatre cens soixante<lb>
quatre<lb>
…
</p>
<pb n="3">
<p>
…A quoy ledict conte de Charroloys, par plusieurs fois,<lb>
voulut respondre comme fort passionné de ceste injure qui<lb>
se disoit de son amy et allyé. Mais ledict Morvillier luy<lb>
rompoit tousjours la parolle, disant ces mots : Monsr de<lb>
Charroloys, je ne suys pas venu pour parler à vous, mais à<lb>
monsr vostre père. Ledict conte supplia par plusieurs foys<lb>
à son père qu'il peust respondre, lequel luy dit : J'ay respondu<lb>
pour toy, comme il me semble que père doit respondre<lb>
pour filz. Toutesfoys, si tu en as si grand envie,<lb>
penses y aujourduy, et demain dy ce que tu vouldras.<lb>
</p>
</body>
</text>
</tei.2>
```

*Un exemple de poésie : extrait d'*Yonec *(*Les lais *de Marie de France)*

```
<!DOCTYPE TEI.2 PUBLIC "-//TEI P4//DTD Main DTD Driver File//EN"
"/dtd/sgml/tei/tei2.dtd" [
<!ENTITY % TEI.general "INCLUDE" >
<!ENTITY % TEI.corpus "INCLUDE" >
<!ENTITY % TEI.analysis "INCLUDE" >
<!ENTITY % TEI.certainty "INCLUDE" >
<!ENTITY % TEI.figures "INCLUDE" >
<!ENTITY % TEI.linking "INCLUDE" >
]>
<!-- En-tête non renseigné -->
<tei.2>
```

```
<teiheader type='text' status='new'>
<filedesc>
<titlestmt>
<title></title>
</titlestmt>
<publicationstmt><p></p></publicationstmt>
<sourcedesc default='no'>
<p><name id='latin' type='lang'></name></p>
</sourcedesc>
</filedesc>
</teiheader>
<text>
<front>
<div type="Prologue">
<head>Prologue</head>
<pb n="1">
<p rend="strophe">
Ki Deus ad duné escïence<lb>
E de parler bone eloquence<lb>
Ne s'en deit taisir ne celer,<lb>
Ainz se deit voluntiers mustrer.<lb>
Quant uns granz biens est mult oïz,<lb>
Dunc a primes est il fluriz,<lb>
E quant loëz est de plusurs,<lb>
Dunc ad espandues ses flurs.<lb>
</p>
…
<pb n="15">
</div>
</front>
<body>
<div type="lai">
<head>Yonec</head>
<p rend="strophe">
Puis que des lais ai comencié,<lb>
Ja n'iert pur mun travail laissié;<lb>
Les aventures que j'en sai,<lb>
Tut par rime les cunterai.<lb>
En pensé ai e en talent<lb>
Que d'Iwenec vus die avant<lb>
Dunt il fu nez, e de sun pere<lb>
Cum il vint primes a sa mere.<lb>
</p>
…
<p type="strophe">
La dame dist qu'ele est malade:<lb>
Del chapelain se prenge garde,<lb>
Sil face tost a li venir,<lb>
Kar grant poür ad de murir.<lb>
La vielle dist:
Vus sufferez !<lb>
Mis sires est el bois alez;<lb>
Nuls n'entrera caënz fors mei.<lb>
Mut fu la dame en grant esfrei;<lb>
Semblant fist qu'ele se pasma.<lb>
Cele le vit, mut s'esmaia;<lb>
L'us de la chambre ad defermé,<lb>
Si ad le prestre demandé,<lb>
E cil i vint cum plus tost pot:<lb>
<foreign>Corpus domini</foreign> aportot.<lb>
Li chevaliers l'ad receü,<lb>
```

```
Le vin del chalice beü.<lb>
Li chapeleins s'en est alez<lb>
E la vielle ad les us fermez.<lb>
</p>
…
</div>
</body>
</text>
</tei.2>
```

Consignes pour le balisage des textes de la base de français médiéval

Serge Heiden (slh@ens-lsh.fr)
Céline Guillot (cguillot@ens-lsh.fr)
Alexei Lavrentiev (alexei-lavrentev@mail.ru)

CNRS / ENS-LSH, FRE 2546

V3.0 2003-02
V2.0 2002-10-30 (intégration du retour des 10 premières relectures)
V1 2002-07

Sommaire

Introduction

Le relecteur/encodeur est chargé à la fois de vérifier que le texte qui lui est confié correspond très exactement à l'édition de référence et d'y introduire un certain nombre de balises XML. Elles se présentent sous la forme de méta-informations qu'il faut insérer entre chevrons[19] dans le texte. On s'efforce d'associer une balise unique à chaque type d'information qu'on veut encoder dans le texte.
On utilise pour cela les balises normalisées de la Text Encoding Initiative (TEI), qui rassemble à ce jour plus de 52 institutions dans 13 pays[20]. Pour encoder les textes de la BFM nous utilisons un sous-ensemble des balises disponibles dans la TEI. Ce sous-ensemble évolue en fonction des idiosyncrasies rencontrées dans les nouveaux textes traités (il serait bien sûr simpliste de penser pouvoir fixer a priori la structure et le contenu de tous les textes). Par ailleurs, l'usage de certaines balises ou de certains attributs est légèrement adapté du standard TEI pour se conformer aux besoins d'exploitation et d'encodage propres à nos textes. Les balises seront la base des traitements automatiques ultérieurs.

Chargement et sauvegarde du fichier

Il suffit d'ouvrir et d'enregistrer le fichier en format texte brut (et non comme un texte HTML comme c'est parfois proposé par défaut). En aucun cas le texte ne doit être transformé en un document Word. Aucun enrichissement typographique ou mise en évidence Word n'est admis.

Principes de base

Fidélité à l'édition de référence

La règle fondamentale de la représentation numérique des textes de la BFM est que la version numérisée doit être la copie exacte du texte papier de l'édition de référence. Toutefois nous avons pour l'instant pris le parti de **ne pas prendre en compte les notes de l'éditeur**.

De cette règle découlent deux conséquences :

- si le relecteur/encodeur désire faire des propositions de correction, il peut le faire au moyen de balises spécialisées à cet effet. **En aucun cas il ne doit modifier directement le texte de l'édition de référence** ;
- nous devons être en mesure de produire à partir de la version numérique du texte une copie fidèle de l'édition de référence (à l'exception des notes). C'est pourquoi les marques typographiques utilisées par l'éditeur scientifique sont une information qui nous est nécessaire (pour une description du balisage de ces marques, se reporter aux sections 6 et 7).

[19] < ... >
[20] Le site web de la TEI se trouve à l'URL http://www.tei-c.org/.

Principes de balisage

- chaque balise est délimitée par des chevrons (**< >**) ;
- il ne doit pas y avoir d'espace entre le nom de la balise et le premier chevron '**<**' ;
- les balises permettent de marquer explicitement la structuration du texte (les divisions de toutes sortes : prologue, chapitre, saut de page…) ;
- on distingue les simples éléments frontières (ex : le saut de ligne marque la frontière entre deux lignes), et les éléments emboîtés qui contiennent d'autres éléments (ex : le chapitre contient une portion du texte et il est susceptible d'être divisé en paragraphes...) :
    - on n'utilise que deux balises du premier type, qui marquent la fin d'une ligne ou d'un vers (**<lb>**[21,22]) et la fin d'une page (**<pb>**[23])
    - dans le second cas, le balisage consiste en un élément ouvrant et un élément fermant qui comporte un slash en préfixe (ex : ** **...** **) ;
- certaines balises comprennent, outre l'élément ouvrant et fermant, un attribut particulier auquel on affecte une valeur (ex : pour délimiter la laisse on utilise la balise <**p>** comportant un attribut **type** encodant le type de division « laisse » **:** ** **…** **).
- l'ordre dans lequel on indique les attributs est libre ;
- le nombre d'espaces ou de tabulations situés entre deux mots ou balises dans le corps du texte n'a pas d'interprétation particulière ;
- de même pour les sauts de ligne.

Délimitation du corps du texte (cf. Section 1 du Manuel d'encodage)

- Si le texte présente un prologue, un appendice ou tout autre élément distinct du corps même du texte, il est nécessaire de distinguer chacun de ces éléments ;
- l'élément **<text>** délimite l'unité textuelle dans sa globalité (y compris le prologue ou l'appendice) ;
- l'élément **<body>** marque le début et la fin du corps du texte proprement dit ;
- l'élément **<front>** enchâsse les informations liminaires : prologue, sommaire…et surtout le titre de l'oeuvre
- l'élément **<back>** enchâsse les informations finales : appendice, index…et surtout l'explicit de l'oeuvre

Les rares explicits ne se trouvant pas en fin de document sont à placer dans une division ** ** dont l'attribut **type** vaut "**explicit**".

La structure potentielle d'un texte est donc toujours la suivante :

```
<text>
<front>
<div type="Prologue">
```

[21] Dans tous les exemples qui suivent, les balises sont indiquées dans une police grasse afin de faciliter la lecture. Dans le corps des textes, elles ne doivent bien sûr porter aucun enrichissement Word particulier.

[22] pour « line-break ».

[23] pour « page-break ».

```
<head>Prologue</head>
<p>
contenu du prologue
</p>
</div>
</front>
<body>
contenu du texte
</body>
<back>
<div type="appendice">
<p>
contenu de l'appendice
</p>
</div>
</back>
<text>
```

ou

```
<text>
<front>
<div type="titre">
<p>
titre de l'oeuvre
</p>
</div>
</front>
<body>
contenu du texte
</body>
<back>
<div type="explicit">
<p>
explicit
</p>
</div>
</back>
<text>
```

Exemple :

```
<text>
<body>
<p>
Buona pulcella fut Eulalia,
Bel auret corps, bellezour anima.
Uoldrent la ueintre li Deo inimi,
Uoldrent la faire diaule seruir.
Elle no.nt eskoltet les mals conselliers,
Qu'elle Deo raneiet chi maent sus en ciel.
…
Tuit oram que por nos degnet preier
```

```
Qued auuisset de nos Christus mercit
Post la mort et a lui nos laist uenir
Par souue clementia.
</p>
</body>
</text>
```

Structure du corps du texte (cf. Section 2 du Manuel d'encodage)

Divisions du texte (cf. Section 2.1 du Manuel d'encodage)

- Le corps du texte (son « body ») peut être structuré en divisions multiples, qui toutes seront indiquées au moyen des balises appropriées ;
- toutes les balises comportent un élément invariant **`div`** ;
- elles ont toutes également un **`type`** qui précise la nature de la division (livre, chapitre, chanson…) et un attribut **`n`** (numéro de chapitre…), ces deux indications étant obligatoirement indiquées (si on les possède).

Exemple :

```
<div type="chapitre" n="1">…</div>
```

```
<div type="livre" n="1">…</div>
```

- si la division comporte un titre, nous vous demandons de l'indiquer au moyen de la balise **`<head>`**.

Exemple :

```
<div type="chapitre" n="1">
<head>Ci commence li premiers chapitres qui parole de l'office as
baillis.</head>
Tout soit il ainsi qu'il n'ait pas en nous toutes les
graces qui doivent estre en homme qui s'entremet de baillie,
pour ce ne lerons nous pas a traitier premiers en cest
chapitre de l'estat et de l'office as baillis, et dirons briement
une partie des vertus qu'il doivent avoir, et comment il se
doivent maintenir, si que cil qui s'entremetront de l'office
i puissent prendre aucune essample...
</div >
```

Pseudo-titres

Il peut arriver qu'on rencontre dans un texte un titre qui ne correspond à aucune division logique (et qui n'est pas numéroté). On considèrera alors qu'on a affaire à une pseudo-

division qu'on balisera au moyen de la balise ** **. Le titre se trouvera comme dans les cas précédents dans son **<head>**.

Changement de manuscrit

Si l'édition de référence marque le changement de manuscrit, il faut l'indiquer dans une note de début et dans une autre note à la fin du changement (cf. section 8 ci-dessous).

```
qui vos devoit faire oblïer ?<lb n='2101'>
<note resp="proofreader">Passage au ms. B</note>Ge vos ai fait
molt lait servise,<lb n='2102'>
car par mon fait estes ocise ;<lb n='2103'>
la sorciere dut enchanter<lb n='2104'>
par coi deüstes oblïer. <lb n='2105'>
<note resp="proofreader">Retour au ms. A</note>
```

NOTA BENE : Le texte de la note est libre.

Numérotation des pages (cf. Section 2.2 du Manuel d'encodage)

Textes en prose

- la pagination est déjà indiquée au moyen de la balise **<pb>** ;
- nous vous demandons de vérifier que la balise se trouve à la bonne place ;
- attention aux illustrations qui peuvent générer des erreurs de pagination, nous vous demandons de nous les signaler (cf. NOTA BENE de la section 2.2 du Manuel).

Textes en vers

- actuellement, nos textes en vers ne comportent pas de numéro de page ;
- nous vous demandons donc d'indiquer le numéro de la page par une balise **<pb>** placée juste avant le premier mot de la page qui commence (les pages seront ensuite numérotées et les balises seront transformées en **<pb n="1">**...) **.**
- attention aux illustrations qui peuvent générer des erreurs de pagination, nous vous demandons de nous les signaler (cf. NOTA BENE de la section 2.2 du Manuel)

Exemple :

```
…
- Or i voist donc, fait ele, car se il demain ne deust
revenir il n'i alast hui par ma volenté.» Et il monte
et la damoisele ausi,
<pb>
si se partent de laienz sanz
autre congié, et sanz plus de compaignie, fors
solement dui escuier qui avec la damoisele
estoient venuz. Et quant il sont issu de Kamaalot
…
```

NOTA BENE : Cas d'un mot qui commence sur une page et finit sur une autre.
Nous vous demandons de supprimer la césure et de mettre le mot tout entier dans la page où il commence

Délimitation des unités inférieures (cf. Section 2.3 du Manuel d'encodage)

Textes en prose

paragraphe

- pour les textes dans lesquels le paragraphe est noté explicitement par la balise ** **…** **, nous vous demandons de vérifier que l'élément ouvrant et l'élément fermant sont à la bonne place (il est inutile de reporter le numéro du paragraphe) ;
- pour les textes dans lesquels les paragraphes sont notés implicitement (par un alinéa ou un saut de ligne) ou ne sont pas indiqués du tout, nous vous demandons de les indiquer de manière explicite au moyen de la balise ** **…** ** ;

ligne

- lorsque les sauts de ligne sont indiqués au moyen des balises <**lb**>, on doit vérifier leur place.

Exemple :

```
<p>
Or i voist donc, fait ele, car se il demain ne deust<lb>
revenir il n'i alast hui par ma volenté.» Et il monte<lb>
<pb>
et la damoisele ausi, si se partent de laienz sanz<lb>
autre congié, et sanz plus de compaignie, fors<lb>
solement dui escuier qui avec la damoisele<lb>
estoient venuz. Et quant il sont issu de Kamaalot<lb>
</p>
```

Textes en vers

strophe, laisse, refrain…

- pour les textes dans lesquels la strophe, la laisse, le refrain… est marqué explicitement par la balise ** **…** **, nous vous demandons de vérifier que l'élément ouvrant et l'élément fermant sont à la bonne place (il est inutile de reporter le numéro de la strophe…) ;

- pour les textes dans lesquels la strophe (ou autre groupe de vers) est notée implicitement (par un alinéa ou un saut de ligne) ou absente totalement du texte, nous vous demandons de l'indiquer de manière explicite au moyen de la balise ** **…** ** (il est inutile d'indiquer l'attribut et le numéro du groupe de vers)

vers

- nous vous demandons de vérifier que la balise **<lb>** se trouve bien à la fin de chaque vers.

Exemple :

```
<text>
<body>
<p rend="gv">
En l'ost n'orent pas lor seignor ;<lb>
en l'andemain matin al jor <lb>
…
que il avoit si tost perdu ; <lb >
molt l'en estoit mal avenu. <lb>
</p>
…
</body>
</text>
```

Numérotation erronée

En cas d'erreur (répétition/suppression) dans la numérotation des vers, il faut l'indiquer dans une **<note>** avant le premier vers concerné (cf. section 8 ci-dessous).

Mises en évidence typographiques (cf. Section 2.3 du Manuel d'encodage)

Conformément au principe énoncé de fidélité à l'édition de référence, nous demandons au relecteur/encodeur de rendre compte des différentes marques typographiques insérées dans le texte par l'éditeur scientifique.

La TEI étant avant tout destinée à représenter les notions sémantiques nécessaires au traitement du texte, nous demandons au relecteur/encodeur de signaler ces marques typographiques au moyen de balises qui en décrivent la signification.

Comme marques typographiques varient considérablement selon les éditions dans leur forme et leur signification, il est donc nécessaire que nous ayons accès à la fois à la signification attachée à chacune de ces marques et à leur forme respective.

Nous demandons donc au relecteur d'analyser l'usage des marques typographiques dans le texte et de remplir le questionnaire *Usage des marques typographiques dans les éditions*

des textes de la BFM qu'on lui fournit. Ce questionnaire donne des indications sur les deux cas de figure qui peuvent se présenter :

- l'usage des marques typographiques est homogène (par exemple, les crochets sont toujours utilisés pour marquer les insertions). Il faut alors utiliser une balise qui décrive la signification de chaque marque typographique (les indications fournies par le questionnaire nous permettant de rétablir les marques typographiques utilisées) ;
- l'usage des marques typographiques n'est pas tout à fait homogène. Il convient alors de préciser la forme de la marque typographique utilisée . On adjoint pour cela l'attribut **`rend`** à la balise qu'on a choisie pour décrire la signification de cette marque.

Les marques de mise en évidence de l'éditeur scientifique sont de deux types :

- marques typographiques au sens propre (italiques, petites majuscules, gras…)
- symboles insérés dans le texte (crochets, parenthèses, points de suspension…)

Le format des textes de la BFM étant le texte brut, il est impératif de baliser conformément à notre usage de la TEI ce que l'éditeur met en évidence dans le texte au moyen des marques typographiques propres (à l'exception des majuscules simples). Les symboles et les majuscules peuvent quant à eux être conservés tels quels dans le texte. Les balises qui encodent leur signification seront par la suite insérées par nos soins grâce aux indications fournies par les réponses au questionnaire.

Corrections du texte (cf. Section 2.3.1 du Manuel d'encodage)

Le système TEI dispose d'un certain nombre de balises pour marquer et gérer les interventions éditoriales aux différentes étapes de la rédaction d'un texte. Pour les textes de la BFM il convient de distinguer en premier lieu les corrections que l'éditeur scientifique apporte à son manuscrit de base et qu'il met en évidence par des marques typographiques, et les corrections des erreurs de l'édition proposées par les relecteurs de la version numérisée.

Corrections de l'éditeur scientifique (cf. Section 2.3.1.1 du Manuel d'encodage)

Dans les cas où le relecteur/encodeur doit encoder des marques typographiques qui correspondent à des corrections de l'éditeur scientifique, il doit utiliser la balise **`<corr>`** pour indiquer que l'éditeur scientifique s'écarte de son manuscrit de base et propose une amélioration du texte (pour le détail du balisage, se reporter à la section 2.3.1.1 du Manuel d'encodage).

NOTA BENE : Si la marque typographique a une autre fonction que la correction (l'italique signale, par exemple, un passage en langue étrangère), il convient d'utiliser la balise prévue à cet effet (cf. section 7 ci-dessous).

Exemples :

```
Lasse, ma proiiere <corr resp="editor">l'a mort</corr> !<lb n="992">
```

Glose : L'éditeur a indiqué en italiques sa correction. La leçon du manuscrit de base (*est la mort*) étant citée en note, on ne la mentionne pas (il n'y a pas d'attribut **sic**). Les italiques ayant toujours la même signification dans ce texte, il n'est pas utile de mentionner leur présence au moyen de l'attribut **rend**.

```
Si a <corr resp="editor" sic="" rend="ital">en</corr> son consel
trouvé
```

Glose : L'éditeur a rajouté une préposition qui était absente du texte d'origine et l'a signalé par des italiques. L'attribut **sic** est présent, mais comme il s'agit d'un ajout sa valeur est nulle. Les crochets étant habituellement utilisés pour marquer les ajouts dans ce texte, il est nécessaire de signaler la présence exceptionnelle des italiques dans ce vers à l'aide de l'attribut **rend**.

```
Sous [c]iel n'a si rice baron, <lb n="523">
```

Glose : Les crochets, qui servent à indiquer l'ajout d'une lettre, sont laissés tels quels par le relecteur dans le texte.

Corrections des relecteurs (cf. Section 2.3.1.2 du Manuel d'encodage)

La règle fondamentale de la représentation numérique des textes de la BFM est que le texte numérisé doit être la copie exacte du texte papier de l'édition de référence. Il est cependant quelques cas exceptionnels où il est possible de mentionner qu'on pense une correction nécessaire. Le passage de l'édition qu'on se propose de corriger est balisé comme élément **<sic>** avec un attribut **corr** contenant la correction proposée. Le contenu de l'élément sera vide en cas d'ajout proposé, et l'attribut **corr** sera vide en cas de proposition de suppression.. L'attribut **resp** permet de faire connaître la personne à l'origine de la correction (le **"proofreader"** désigne le relecteur et le **"encoder"** désigne la personne responsable de la fin du balisage et de la vérification finale).
Deux cas de figure se présentent au relecteur :

1er cas :
Une intervention éditoriale est nécessaire pour que le texte numérisé corresponde à la version de référence (le texte numérisé est différent du texte papier) et le relecteur/encodeur pense que l'éditeur de la version numérique (par opposition à l'éditeur de la version papier) avait une intention de correction. Il décide donc de conserver cette correction et en prend la responsabilité.

Exemple :
que **<sic resp="proofreader" corr="">**que**</sic>**
ce qui signifie que la version numérique corrige une répétition (qui fausse le vers par exemple)

2ème cas :
Le relecteur veut proposer lui même une correction.

Exemple :
La typographie d'une édition de référence ne différentie par le I du L. Si le correcteur estime pour des raisons linguistiques que la graphie « I' » doit être résolue en « L' » à certains endroits du texte, il l'indique par : **<sic resp="proofreader" corr="L">**I**</sic>**'.

Partie manquante (cf. Section 2.3.1.3 du Manuel d'encodage)

Les parties manquantes de manuscrits sont à indiquer à l'aide d'un élément **<gap>**. Cette information est donnée par l'édition de référence. Le responsable de la balise est donc l'éditeur scientifique (**resp="editor"**).

Exemple :

```
Si angoissés <gap resp="editor" rend="points"><lb n="4887">
Que je ne pooie dormir,<lb n="4888">
```

Glose : Il manque ici, d'après l'éditeur scientifique la fin du vers.

Délimitation des passages en langue étrangère (cf. section 2.3.2.1 du Manuel d'encodage)

Si vous rencontrez un passage en langue étrangère (en latin le plus souvent), nous vous demandons de l'enchâsser dans le texte au moyen de la balise **<foreign>**…**</foreign>**. S'il s'agit d'une autre langue que le latin, il convient de l'indiquer dans l'attribut **lang**. Si l'usage de marques typographiques qui servent à mettre en évidence les passages en langue étrangère n'est pas systématique, il convient de préciser la marque typographique utilisée dans chaque cas à l'aide d'un attribut **rend**. Si le passage n'est pas marqué typographiquement, il faut utiliser **rend="non"**.

Exemple :

```
…
- Et tout li haut
homme, et clerc et lai et petit et grant, demenerent
si grant goie a l'esmovoir que onques encore si faite
goie ne si fais estoires ne fu veus ne oïs; et si fisent
li pelerin monter as castiaus des nes tous les
prestres et les clers qui canterent <foreign rend="ital"> Veni
creator spiritus </foreign>
Et trestout et grant et petit plorerent de pec et de
le grant goie qu'i eurent
…
```

Autres mises en évidence (cf. Section 2.3.2.2 du Manuel d'encodage)

Dans le cas où une marque typographique a un autre usage que ceux qui viennent d'être décrits, nous vous demandons de baliser le passage mis en évidence au moyen de l'élément **<hi>** et d'ajouter une **<note>** après le passage balisé, qui décrive si possible la fonction de cette mise en évidence typographique.
Les valeurs proposées dans la colonne intitulée "fonction" du questionnaire *Usage des marques typographiques dans les éditions des textes de la BFM* peuvent guider la rédaction de cette note.

Exemples :

texte de l'édition	texte balisé
si com Tulles le determine ou livre qu'il fist *de Viellece*, qu'il loe et veust plus que Jeunece,	si com Tulles le determine**<lb>** ou livre qu'il fist **<hi>**de Viellece**</hi><note>**titre d'une autre œuvre**</note>,<lb>** qu'il loe et veust plus que Jeunece,**<lb>**

texte de l'édition	texte balisé
Bele, vers cui mes cuers s'acline, RENALS DE BIAUJU molt vos prie Por Diu que ne l'oblïés mie.	Bele, vers cui mes cuers s'acline,**<lb>** **<hi>**Renals de Biauju**</hi><note>**nom de l'auteur</note>molt vos prie**<lb>** **<pb>** Por Diu que ne l'oblïés mie.**<lb>**

texte de l'édition	texte balisé
en l'an de Nostre Seignor Dieu Jhesu Crist a mille .iij^{c}x. ans, le .xv^{e}. jour dou mois de mars par .j. leundi de la .viij^{e}. indicion.	en l'an de Nostre Seignor Dieu Jhesu Crist a mille .iij**<hi>**c**</hi><note>**chiffre en exposant**</note>** x. ans, le .xv**<hi>**e**</hi><note>**chiffre en exposant**</note>**. jour dou mois de mars par .j. leundi de la .viij**<hi>**e**</hi><note>**chiffre en exposant**</note>**. indicion.**</p>**

Notes ou commentaires (cf. Section 2.3.3 du Manuel d'encodage)

L'élément **<note>** est un élément réservé à l'usage du relecteur/encodeur. En effet, il lui permet d'insérer des commentaires dans n'importe quel endroit du texte. Ces commentaires seront alors lisibles par les relecteurs finaux tout en étant clairement marqués comme distincts du texte. En cas de doute lors de l'interprétation d'un passage du texte, nous demandons au relecteur de toujours utiliser un élément **<note>** pour exprimer sa remarque, à l'endroit même où le problème se pose. Il est tout à fait normal de ne pas pouvoir trancher dans certaines situations et toute décision ou non décision doit toujours être explicitement marquée dans les textes. Chaque note doit être "signée" à l'aide de l'attribut **resp**.

Exemple :

```
A Com cist cheualiers qui ci siet. <note resp="proofreader">Deux
lettres majuscules initiales</note> <lb n="423">
Qu'il ne respont ne un neel.
```

Glose : l'éditeur indique la présence de 2 majuscules en début de ligne.

Questionnaire

Usage des marques typographiques dans les éditions des textes de la BFM

Nom du relecteur _____
Titre de l'oeuvre_____
Editeur scientifique _____
Collection _____ Année _____

Ce questionnaire doit nous permettre à la fois de rétablir les marques typographiques utilisées dans le texte par l'éditeur scientifique et d'encoder la fonction qu'elles remplissent.

Notre objectif étant par ailleurs de simplifier au maximum le travail de nos relecteurs, nous leur demandons d'encoder dans le corps du texte uniquement la fonction attachée aux marques typographiques (par exemple, l'italique est utilisé pour marquer les mots latins), ce travail étant celui qui requiert une expertise en ancien français et en matière d'édition critique. La forme de ces marques pourra être rétablie par nos soins grâce à nos outils informatiques si une fonction est toujours marquée typographiquement de la même manière (par exemple, tous les mots latins du texte sont en italiques). Si tel est le cas, nous considérons que l'usage de cette marque est systématique (cf. 3ème colonne du tableau ci-dessous).

Une même marque typographique peut être systématiquement employée pour plus d'une fonction (par exemple, l'italique est utilisé pour marquer les mots latins et les titres d'œuvre), ce qui implique de cocher plusieurs cases pour une même marque dans le tableau.

Si le marquage d'une fonction n'est pas systématique (par exemple, la correction d'un mot du manuscrit est indiquée tantôt au moyen de crochets, tantôt au moyen d'italiques, ou bien les mots latins du texte sont tantôt mis en italiques, tantôt non marqués), on dira qu'il est occasionnel (cf. 4ème colonne du tableau). Dans ce dernier cas, il est nécessaire d'indiquer dans le corps du texte quelle est la marque typographique utilisée (pour une description du balisage de ces marques, se reporter aux *Consignes de balisage*).

Dans le tableau qui suit sont décrites un certain nombre de fonctions communément associées aux marques typographiques. Si dans une édition une marque typographique a une fonction qui n'apparaît pas dans cette liste, il est possible de la rajouter dans la case "autre".

Marque	Fonction	Usage systématique	Usage occasionnel	Non utilisé
Italique	Passage dans une langue étrangère	☐	☐	☐
	Titre d'une autre oeuvre	☐	☐	☐
	Citation à l'intérieur du discours direct	☐	☐	☐
	Autre type de citation (préciser)	☐	☐	☐
	Correction (changement de ms.)	☐	☐	☐
	Correction (modification)	☐	☐	☐
	Correction (ajout)	☐	☐	☐
	Résolution d'abréviation	☐	☐	☐
	Autre (préciser) _____	☐	☐	☐
Gras	Titre d'une autre oeuvre	☐	☐	☐
	Citation	☐	☐	☐
	Correction	☐	☐	☐
	Autre (préciser) _____	☐	☐	☐
MAJUSCULES	Titre de l'oeuvre	☐	☐	☐
	Titre d'une autre oeuvre	☐	☐	☐
	Premier mot du volume ou d'une autre division du texte	☐	☐	☐
	Nom de l'auteur	☐	☐	☐
	Citation	☐	☐	☐
	Chiffres romains	☐	☐	☐
	Autre (préciser) _____	☐	☐	☐
PETITES MAJUSCULES	Titre de l'oeuvre	☐	☐	☐
	Titre d'une autre oeuvre	☐	☐	☐
	Premier mot du volume ou d'une autre division du texte	☐	☐	☐
	Nom de l'auteur	☐	☐	☐
	Citation	☐	☐	☐
	Chiffres romains	☐	☐	☐
	Autre (préciser) _____	☐	☐	☐
Exposant ^{x}	Chiffre en exposant	☐	☐	☐
Crochets [x]	Titre de l'oeuvre	☐	☐	☐
	Correction (ajout)	☐	☐	☐
	Numéro de folio, etc.	☐	☐	☐
	Autre (préciser) _____	☐	☐	☐
Crochets avec points de suspension […]	Lacune ou omission	☐	☐	☐
	Autre (préciser) _____	☐	☐	☐
Points de suspension …	Lacune ou omission	☐	☐	☐
	Autre (préciser) _____	☐	☐	☐
Points (plus de 3)……	Lacune ou omission	☐	☐	☐
	Autre (préciser) _____	☐	☐	☐
Parenthèses	Ponctuation courante	☐	☐	☐
	Correction (ajout)	☐	☐	☐
	Autre (préciser) _____	☐	☐	☐
Autre _____		☐	☐	☐
_____		☐	☐	☐

Liste d'autorité pour les valeurs de l'attribut rend des éléments suivants : <corr>, <hi>, <gap> et <foreign>[24]

Valeur	Description	Exemple
ital	italiques	*creator*
gras	gras	**creator**
maj	majuscules	ENEAS
pmaj	petites majuscules	RENAUD DE BEAUJEU
exp	exposant	XX^{C}
ind	indice	XX$_{C}$
crochets[25]	cas où il y a plusieurs mots à l'intérieur des crochets	[et nen estoit leus de deffendre. Tote ert la vile mise en cendre].
susp	3 points de suspension	…
points	plus de 3 points de suspension	……
crochets-susp	3 points de suspension entre crochets	[…]

NOTA BENE

En cas de correction de l'éditeur scientifique qui rajoute soit un mot, soit un ou plusieurs caractère(s) à l'intérieur d'un mot, et qui le note dans l'édition au moyen de crochets, on utilise l'élément <corr> dont l'attribut rend restitue la chaîne de caractères telle qu'elle se présente dans l'édition.

Exemple :

```
<corr resp="editor" sic="travalle" rend="trava[i]lle">travaille</corr>
```

Pour des raisons techniques il serait difficile d'utiliser la valeur habituelle "crochets" dans ce cas. Si nous le faisions, nous serions soit conduits à éclater le mot en plusieurs unités, ce qui pourrait perturber l'exploitation automatique ultérieure de la base, soit incapables de repérer la position des crochets situés à l'intérieur d'un mot.
Par extension et pour faciliter le balisage automatique des textes de notre base, ce principe de notation a été adopté dans les cas où les crochets enserrent un seul mot.

[24] A distinguer de la liste d'autorité des valeurs de l'attribut rend de l'élément <p>.
[25] voir le NOTA BENE.

Liste d'autorité pour les valeurs de l'attribut rend de l'élément <p> pour les textes en vers [26]

Valeur	Description
strophe	strophe
laisse	laisse
couplet	couplet
gv	groupe de vers sans dénomination particulière

De façon générale, la valeur sera déterminée à partir du terme qu'utilise l'éditeur scientifique. En cas d'absence, on emploiera la valeur "gv".

[26] Pour les textes en prose, l'attribut rend n'est pas utilisé avec l'élément <p>.